Journal of Information Systems Engineering and Management
2024, 9(4), 30195
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Improving Seasonal Influenza Forecasting Using Time Series Machine Learning Techniques

Salem Mubarak Alzahrani [1], Fathelrhman EL Guma [1,2*]

[1] Doctor, Faculty of Science, Al-Baha University, Al Baha, Saudi Arabia
[2] Doctor, Department of Statistics and Population Studies, Alsalam University, Alfula, Sudan
* **Corresponding Author:** fathyelrhman@gmail.com

Citation: Alzahrani, S. M., & Guma, F. E. (2024). Improving Seasonal Influenza Forecasting Using Time Series Machine Learning Techniques. *Journal of Information Systems Engineering and Management*, 9(4), 30195. https://doi.org/10.55267/iadt.07.15132

ARTICLE INFO

Received: 09 Aug 2024
Accepted: 27 Aug 2024

ABSTRACT

Influenza is a highly contagious respiratory disease and is still a serious threat to public health all over the world. Forecasting techniques help in monitoring seasonal influenza and other influenza-like diseases and also in managing resources appropriately to formulate vaccination strategies and choose appropriate public health measures to reduce the impact of the disease. The aim of this investigation is to forecast the monthly incidence of seasonal flu in Saudi Arabia for the years 2020 and 2021 using the XGBoost model and compare it with ARIMA and SARIMA models. The results show that the XGBoost model has the lowest values MAE, MAE, and RMSE compared to the ARIMA and SARIMA models and the highest value of R-squared (R^2). This study compares the accuracy of the XGBoost model with ARIMA and SARIMA models in providing a forecast of the number of monthly seasonal influenza cases. These results confirm the notion that the XGBoost model has a higher accuracy of prediction than that of the ARIMA and SARIMA models, mainly due to its capacity to capture complex nonlinear relationships. Therefore, the XGBoost model could predict monthly occurrences of seasonal influenza cases in Saudi Arabia.

Keywords: Seasonal Influenza, ARIMA, SARIMA, XGBoost, Machine Learning, Public Health, Prediction.

INTRODUCTION

Seasonal influenza, or “the flu”, is a lung disease caused by different types of influenza viruses that affect people worldwide every year. Devlin (2008) states that Influenza usually presents with an upper respiratory illness. The various symptoms of the disease are fever, stuffy nose, sore throat, cough, headache, fatigue, and muscle pain. However, influenza can sometimes lead to severe or fatal pneumonia from the virus or a bacterial illness in the lower respiratory tract that follows (Nelson & Holmes, 2007).

Four types of influenza viruses are available, including A, B, C, and D. Influenza A is the most prevalent and, along with type B, is responsible for the annual outbreaks in humans. Some environmental factors influence these outbreaks, as the virus can mutate rapidly. It can avoid host defences and perhaps raises the infectivity and virulence of some epidemics (Dandachi et al., 2024; Alsubaie, EL Guma, Boulehmi, Al-kuleab, & Abdoon, 2024; Badar et al., 2024). Because of these mutations, it becomes vital that efforts used in controlling the flu virus be seasonal so that proper vaccines can be developed (Dandachi et al., 2024).

Casting the prevalence and emergence of infectious diseases, including influenza into the future is a crucial component of long-term health planning and short-term treatment intervention. To this end, a number of statistical models and machine learning have been utilised for disease prevalence prediction. These models

Copyright © 2024 by Author/s and Licensed by IADITI. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

comprise the seasonal autoregressive integrated moving average (SARIMA) models which are particularly useful in capturing the seasonal variation (Lv, An, Qiao, & Wu, 2021; Nelson, 1998; Mills, 2019), support vector machines (Alzahrani, 2024) and the extreme gradient boosting (XGBoost) models (EL Guma, 2024; Song, 2017). Though, generalized additive models and deep learning models including the neural network model in the long short-term memory have been used for the epidemic forecasting to capture linear effects, while ARIMA has been more efficient for capturing the linear, seasonal effect. However, it cannot represent the non-linear feature of the influenza outbreak (Bezerra & Santos, 2020). On the other hand, the XGBoost has the advantage of better performance on large datasets. Besides, it can clearly express nonlinear characteristics; and it is a promising model for enhancing the precision of predictions (Aljandali, 2017; Kumar, Thiruvarangan, Vishnu, Devi, & Kavitha, 2022; Peixeiro, 2022; Q. Chen, 2024; Luo, Zhang, Fu, & Rao, 2021). Therefore, the current research aims at creating and evaluating an XGBoost model that estimates the subsequent month's influenza cases in Saudi Arabia in contrast to the ARIMA and SARIMA models. By comparing the predictions according to all these models, we aim to develop a solid basis to predict the occurrences of seasonal influenza cases. In turn, it helps to manage, stabilise and predict the annual epidemics of influenza in Saudi Arabia.

LITERATURE REVIEW

Forecasting especially the anticipatory seasonal influenza outbreaks is vital for purposes of both public health and vaccinations. A great deal of literature focuses on the issues related to predicting epidemic diseases, especially the flu viruses of the seasonal kind. These studies have looked at rate of occurrence and mutation within the virus, and the effect of environmental factors on the breakout (Nelson & Holmes, 2007). Disease forecasts that include the accurate infection risk and its time of occurrence allow faster treatment and reduce disease incidence (Tenepalli & TM, 2024). Most of the subsequent research has employed conventional statistical models, including ARIMA and SARIMA, on confirmed cases and death time series concerning epidemic diseases. Influenza forecasting was investigated by Arwaekaji, Sillabutra, Viwatwongkasem and Soontornpipit (2022) in a work that used a SARIMA model with a seasonal autoregressive moving average of the Box-Jenkins. Taking monthly influenza virus infections in Public Health Region 8 Udonthani, Thailand from January 2016 to December 2018 the competing models were compared to identify the best based on AIC, BIC, and RMSE. SARIMA (1,0,1)(1,0,0)12 was predicted to be the best model of influenza forecasting based on the result of this study. It also yielded the least value of AIC (59. 24), BIC (67. 16) and RMSE (0. 4574). Similarly, the actual and the predicted values of the tests were evaluated, and the mean absolute percentage error was 24. 15%. This serves to confirm that the model should be able to predict the existence of influenza and not only that, to provide empirical evidence thereof.

In another study, using SARIMA models in conjunction with seasonal regression integrated moving average, Y. Chen et al. (2020) attempted to estimate the ILI% in Shenyang. The purpose of this work was to provide the epidemiological and causal profiles of the influenza disease. Data from influenza surveillance of the proportion positive for the influenza virus and incidence of ILI from 18 countries over the period 2010-2019 were used for the paper. For forecasting the incidence rates of Influenza for the urban areas SARIMA (0, 1, 0) (0, 1, 2)12 was significant while for the rural area SARIMA (1, 1, 1) (1, 1, 0)12 was also significant. They established the pattern of regional and seasonal distribution of ILI activity in Shenyang and their results confirmed the study. Moreover, the influenza activity above described pointed to the general epidemiological profile of the circulating strains of influenza virus. They also realized that infection by the influenza virus was more common in these small infants rather than in adults.

To forecast the epidemiological trends of the COVID-19 pandemic in several countries, Arun Kumar et al. (2021) utilized time series models such as Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA) and Autoregressive Integrated Moving Average (ARIMA). The model employed was ARIMA and the parameters of the model were estimated following a good fitness of the forecasted outcomes with the test data. ACF, PACF for the models' performance and AIC and BIC for model selection were used. However, other criteria such as the mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and mean relative absolute error (MAPE) shall be used in arriving at which model is most appropriate. The outcomes witnessed the predicted patterns of confirmed instances, recoveries, and total deaths and confirmed an augmentation in some nations concerning these factors. Moreover, forecasting carried out in this paper using the SARIMA model was better than the one using the ARIMA model indicating the presence of seasonality in the COVID-19 data.

To predict influenza in Shanxi Province, Zhao, et al. (2023) developed a SARIMA model for influenza, SARIMA-LSTM combined model, SSA-SARIMA-LSTM combined model. For the first week of 2010 to the 52nd

week of 2019, information on the epidemic's seasonal characteristics was obtained using the seasonal trend decomposition technique. Three other validity measures namely MSE, MAE and RMSE were also used to check the efficiency of the model. Such a pattern of the epidemic time series was observed in seasonal variations, where the value declined from year to year. The sickness was at its worst towards the end of the year and towards the beginning of the following year. The SARIMA-LSTM model excelled. As compared to the SARIMA model it had given lesser mean square error, mean absolute error and mean square error.

EL Guma (2024) compared SARIMA and LSTM models to predict visceral leishmaniasis in eastern Sudan. The research data were collected from health offices from January 2000 to December 2021. The performance of each model was assessed using three metrics: mean average accuracy (MAP), root mean square (RMS), and mean absolute error (MA). The findings indicate that the LSTM model surpasses the SARIMA model regarding prediction accuracy.

METHODOLOGY

Classical Time Series Models

Autoregressive (AR) Integrated Moving Average (MA) Model

The ARIMA model is derived by integrating the AR(p) and MA(q) after adjusting for the essential differences to ensure the time series is stationary. This model is called ARIMA (p, d, q) (Box, 2013).

$$X_t = \Delta_0 + \Delta_1 X_{t-1} + \Delta_2 X_{t-2} + \ldots + \Delta_p X_{t-p} + \epsilon_t - \Theta_1 \epsilon_{t-1} - \Theta_2 \epsilon_{t-2} - \ldots - \Theta_q \epsilon_{t-q} \quad (1)$$

Where

$\Delta_0, \Delta_1, \ldots\ldots, \Delta_p$ are the autoregressive parameters, $\Theta_1, \Theta_2, \ldots, \Theta_q$ are the moving average parameters, and ϵ_t is the error term at time t.

When the time series exhibits non-stationarity, it may be converted into a stationary series by using differencing of order (d). The ARIMA model is represented as ARIMA (p, d, q) in this scenario (Kaur, Parmar, & Singh, 2023), with its equation expressed as:

$$\Delta_p(B)\nu^d X_t = \Theta_q(B)\epsilon_t \quad (2)$$

Where d is the differencing order to achieve stationarity in the time series, and the backward shift operator is denoted by B.

Seasonal AR Integrated MA (SARIMA (p, d, q) (P, D, Q))

The SARIMA model consists of the non-seasonal ARIMA(p, d, q) model with extra seasonal terms (P, D, Q)S. These seasonal terms account for the seasonality of the time-series data, where S is the number of time steps corresponding to a single seasonal period. P, Q, and D represent the order of the seasonal AR term, MA term, and seasonal differencing term, respectively (Anderson, 1977; Khan, Patankar, & Khan, 2024). The mathematical representation of the SARIMA model is as follows:

$$\Delta_P(B)\Lambda_p(1 - B^S)^D(1 - B)^d X_t = \Theta_Q(B^S)\Theta_q(B)\epsilon_t \quad (3)$$

Where X_t represents the non-stationary time series, ϵ_t represents the Gaussian white noise process, $\Delta_P(B)$ represents the non-seasonal AR polynomial, and $\Theta_q(B)$ represents the non-seasonal MA polynomial, and D represents the seasonal differencing component.

Box-Jenkins Approach

The Box-Jenkins approach consists of four essential stages for determining the most suitable model to predict the time series under investigation (Yasmin & Moniruzzaman, 2024; Zhang, Bian, Qu, Tuo, & Wang, 2021; Yenilmez & Mugenzi, 2023).

Identification Stage

This stage aims to identify the model parameters p, d, and q. The Dickey-Fuller tests select the required series of differences d until the time series becomes stationary. The autocorrelation function (ACF) and the partial autocorrelation function (PACF) are employed to determine the values of p and q. Some criteria, including the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), are employed to compare models and select the most suitable one.

Estimation Stage

After the proposed model that represents time series data is identified, the model parameters are estimated.

The most significant parameter estimation method is the Maximum Likelihood.

Diagnostic Stage

For Diagnostic, the proposed model represents the time series data from the previous phases. The model's suitability for the data is assessed by analyzing the residuals that result from its application.

Forecasting Stage

The forecasting stage is the last step of the Box-Jenkins model and the major purpose of the model development strategy.

XGBoost Mode

XGBoost model was introduced by Shen Tianqi and Carlos Gestrin (2011) to improve prediction accuracy. The model has become famous for building predictive models as large or complex datasets can be utilized (EL Guma, Musa, Alkhathami, Saadeh, & Qazza, 2023; Lv, An, Qiao, & Wu, 2021). Through this model, weak prediction models are combined into a unified model for improving forecasting accuracy. Therefore, XGBoost uses a second-order Taylor expansion of the cost function and a regularization term to prevent overfitting. Both the first-order and second-order derivatives are leveraged to compute the pseudo-residuals necessary for learning (Khan et al., 2024). The loss function in XGBoost can be expressed as:

$$y^k = \sum_{i=1}^{n} \iota\left(\left(y_i, y^k{}_i\right) + f_k(x_i)\right) + \Omega(f_k) \quad (4)$$

When XGBoost is executed, learning feature scores build a decision tree. Via these scores, more accurate predictions can be expected. The algorithm is efficient in that a default direction for handling missing or sparse values is specified to optimize the identification of segmentation points (Yasmin & Moniruzzaman, 2024). As classical models struggle with complex data, the XGBoost method is exclusively chosen for the current study as large-scale data sets can be handled and managed with high predictable accuracy and computational efficiency (Sroka, 2024).

Hyperparameter Tuning

The classical method for hyperparameter tuning is grid search, which is an exhaustive search of a given subset of possible values in hyperparameter space. A grid search algorithm is guided by a performance measure, often derived by cross-validation on the training set. In addition, scholars have suggested other tuning methods, such as random search (Yasmin & Moniruzzaman, 2024), gradient search (Yenilmez & Mugenzi, 2023), and Bayesian optimization (Li, Yin, Quan, & Zhang, 2019; EL Guma et al., 2023; Man, Huang, Qin, & Li, 2023; Hoque & Aljamaan, 2021). Random Search randomly selects hyperparameter combinations instead of exhaustively searching through all potential possibilities in a grid search. This is relevant to discrete and continuous scenarios—hyperparameters with mixed values.

Grid search was the primary method used for optimization in this study, for loops in Python were used to go through a grid of predefined hyperparameter values. The machine learning model was trained on a new set of hyperparameters after each pass, and cross-validation was used to test how well it did. The grid included learning rates, regularization strengths, and tree levels in the hyperparameter space. With the for loops, it was possible to study the model's behavior across the grid in great detail, which led to the best performance measure. Even though it was hard to compute, the grid search method made it easy to make hyperparameter changes that were open and clear, which helped me choose the best model setup for the project.

Evaluations Criteria

The metrics commonly employed to assess the accuracy of the proposed model are MAE, MSE, RMSE, MAPE, and coefficient of determination (R^2) (Alsobhi, 2022; Dancer & Tremayne, 2005; Kuran, Tanırcan, & Pashaei, 2023). The metrics are given as:

$$\text{MAE} = \frac{\sum_{i=1}^{T} |X_I - \hat{X}|}{T} \quad (5)$$

where T is the number of errors, and $\sum_{i=1}^{T} |X_I - \hat{X}|$ is the absolute errors.

MSE calculates the squared distances between data points and a regression line to determine its closeness. Squares are needed to eliminate negative values given by Equation (6):

$$\text{MSE} = \frac{\sum_{i=1}^{n} (\chi_i - \widehat{\chi_i})^2}{T} \quad (6)$$

The RMSE is mathematically described by Equation (7) and quantifies the standard deviation of the prediction errors.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\chi_i - \widehat{\chi_i})^2}{n}} \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{T}(\chi_i - \hat{\chi}_i)^2}{\sum_{i=1}^{T}(\chi_i - \bar{\chi}_i)^2} \quad (8)$$

The symbol $\hat{\chi}_i$ denotes the predicted value of the i^{th} sample, whereas T represents the total number of samples and represents the corresponding actual value.

RESULTS AND DISCUSSION

Data Collection

We utilized a dataset encompassing monthly seasonal influenza cases from January 2017 to December 2022. The data was obtained from the offices of the Ministry of Health in the Kingdom of Saudi Arabia (WHO Influenza Surveillance Program, 2008).

Data Splitting

To normalize the process of developing and evaluating algorithms, we divided this dataset into significant components for our experiment on time series forecasting. The data was partitioned into subgroups for both training and testing. Approximately 84% of the dataset was included in the training set, which was used for parameter optimization and model training. The remaining 16% was allocated to the test group. Setting up Hyperparameters.

Features of Monthly Seasonal Influenza Cases

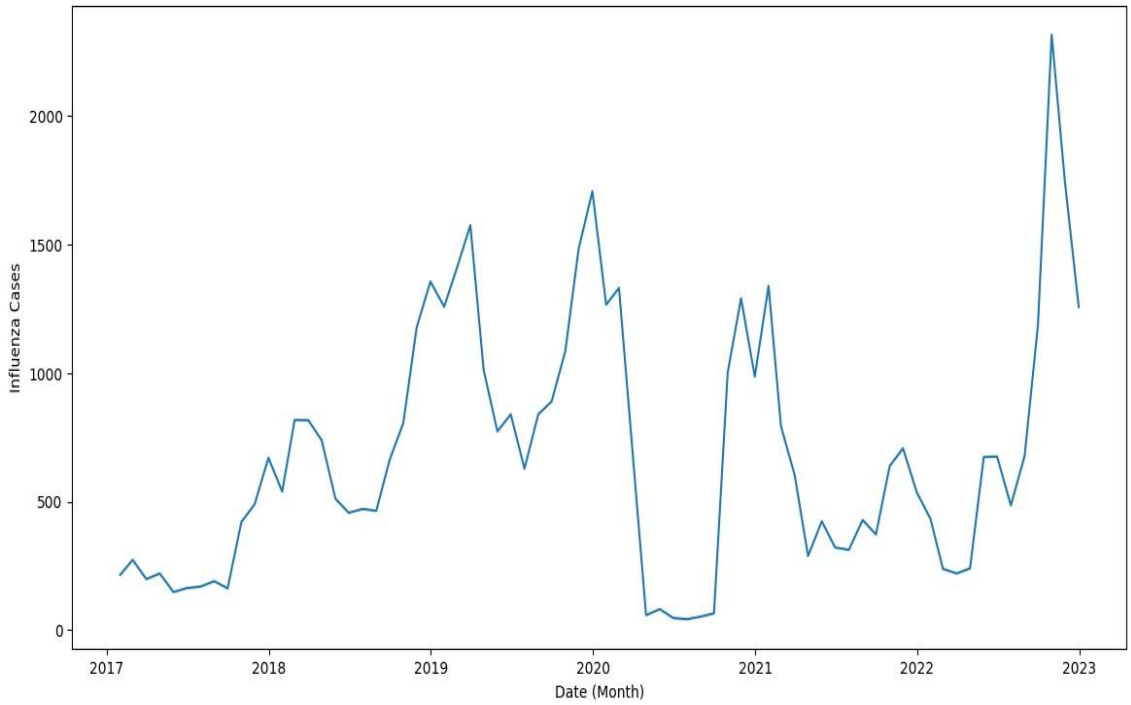

Figure 1. Monthly Seasonal Influenza Cases in Saudi Arabia from January 2017 to December 2022

Figure 1 illustrates the variations in the prevalence of influenza cases across the years. Fluctuations in influenza transmission occur at certain times, with notable spikes and declines, suggesting seasonal variability. The time series plots indicate a noticeable decrease in the early months of 2017. Subsequently, there was a notable surge in cases at the end of 2017 and the beginning of 2018, trailed by a substantial decrease in initial 2020, corresponding with the onset of the COVID-19 pandemic. The decrease is ascribed to interventions such as social separation. Enforced restrictions on movement and heightened sanitation measures. Following a substantial decrease in 2020, there is a steady rise in influenza incidence from 2021 to 2023. This suggests a resumption of regular life as limitations associated with the epidemic are gradually lifted. The data generally suggest a recurring pattern that aligns with the influenza seasons, characterized by high and low points. Moreover, the COVID-19 pandemic has substantially influenced the spread of influenza, resulting in a noticeable decrease in occurrences during 2020. The increase in influenza cases during the post-pandemic era of 2022-2023 emphasizes the need to

monitor and track influenza outbreaks after relaxed pandemic restrictions.

Figure 2 (top part) illustrates the association between a time series and its previous values in the ACF plot. The initial latencies (0 to 5) demonstrate a significant degree of autocorrelation, which implies a robust connection between past and future values within these delays. Progressive declines in the autocorrelation pattern suggest a gradual decline. The observation of significant deviations from the confidence interval confirms the presence of a non-random pattern or seasonality in the data.

Figure 2 (bottom part) illustrates the PACF diagram, which shows the correlation between a time series and its own lagged values while also considering the impact of previous delays. The initial latency significantly increases, indicating a robust positive correlation with the value immediately preceding it.

During the initial five intervals, activity significantly increases following the initial latency. The correlation decreases at a higher rate than the ACF, suggesting that the initial delays may account for the preponderance of the correlation.

Figure 3 illustrates influenza's seasonal decomposition. The graphic depicts influenza's raw time series data. The existence of discernible peaks and troughs implies a recurring pattern of surges and declines in instances.

The trend component represents the fundamental, enduring movement in the series over an extended period. The data indicates a gradual increase until approximately 2019-2020 when it experienced a decline and a subsequent surge in current. The seasonal component represents the recurring short-term pattern in the series. A robust seasonal influence is indicated by the pattern's high level of consistency on an annual basis, characterized by frequent peaks and troughs.

The residual component is the series portion that remains after the trend and seasonal components have been eliminated. The data oscillates around zero due to random noise or inexplicable volatility. The Augmented Dickey-Fuller (ADF) test verified the stability of the fitting analysis on the training dataset of influenza data in this investigation with a significance level of $p<0.001$.

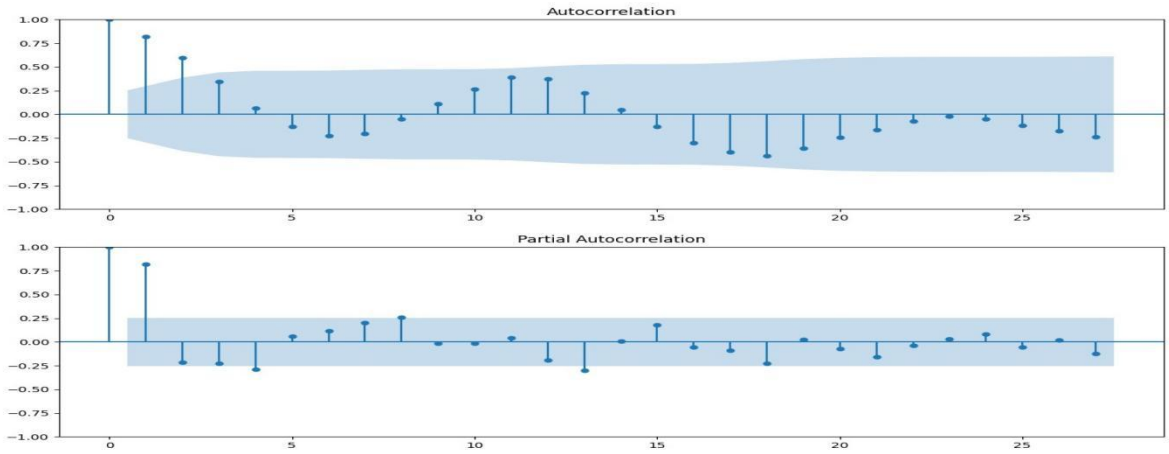

Figure 2. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of Original Monthly Influenza Cases for the Training Data Set

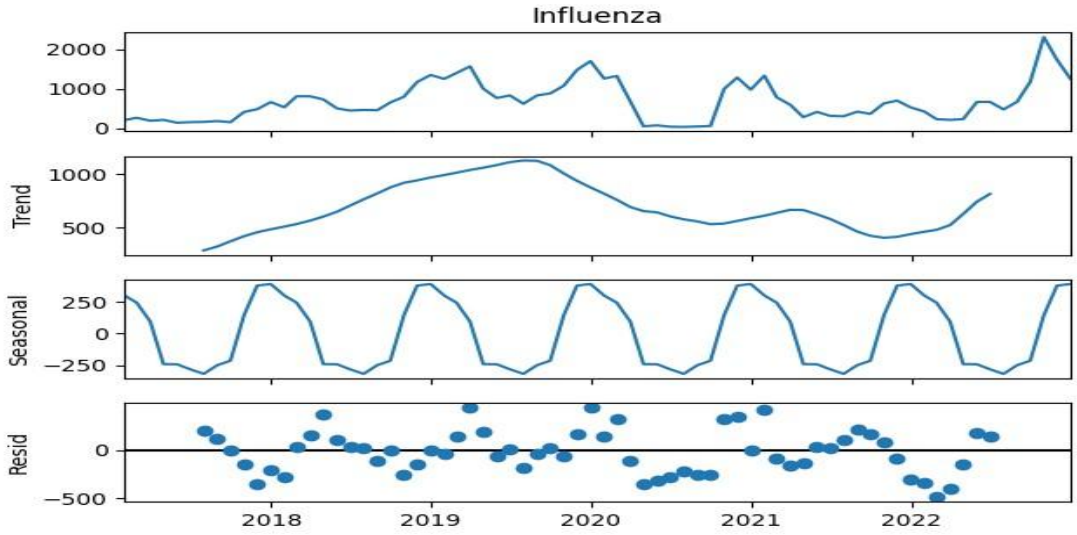

Figure 3. The Seasonal Decomposition of Monthly Influenza Cases in Training Data Set

Forecasting Influenza Monthly Cases Applying the ARIMA Model

we utilized the grid search method for tuning the hyperparameters to identify the most optimal ARIMA model with the lowest AIC number. With an AIC value of and a BIC value, ARIMA(3, 0, 2) is the best model.

The figure compares the real monthly influenza case data with the forecasts generated by the ARIMA model. The model accurately represents the general pattern and notable fluctuations in flu cases. While there are significant disparities between the projected numbers and the real data, particularly between the high and low points of the illness. The outcomes derived from the ARIMA model are consistently similar to the main trend of the real data. The ARIMA model correctly shows how the number of seasonal influenza cases increased towards the year's end. The model usually fits the big picture, but sometimes it either overestimates the real number of cases, especially when a lot is going on. Seasonality: Some of the changes that have been seen might be because the model doesn't take seasonal factors into account.

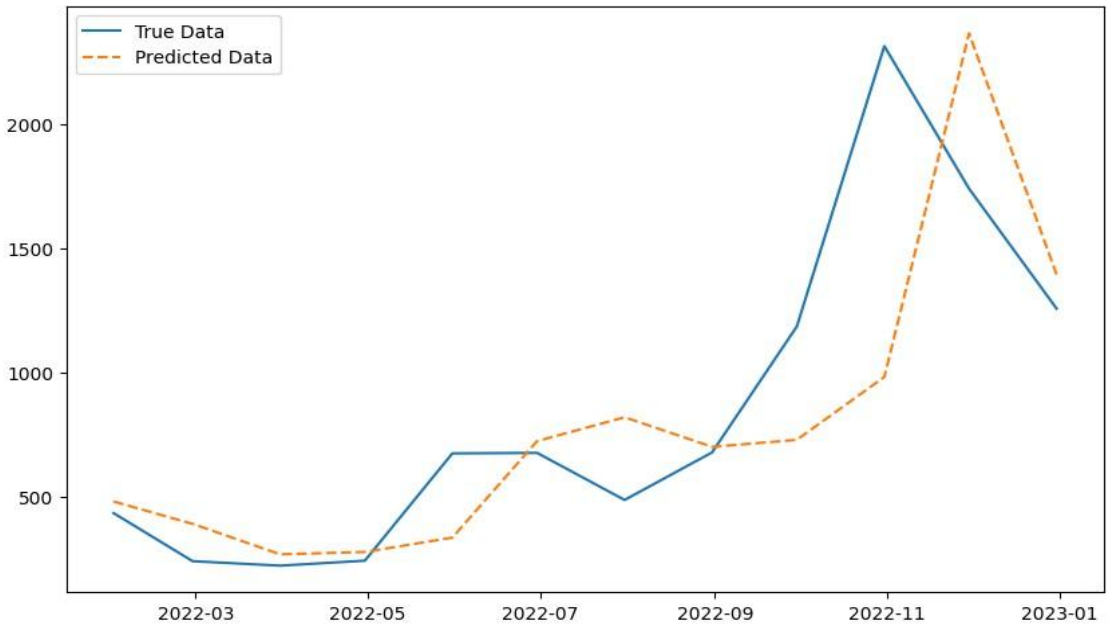

Figure 4. Comparative Evaluation of Monthly Influenza Cases Versus ARIMA Predictions

Forecasting Influenza Monthly Cases Applying the SARIMA Model

The Augmented Dickey-Fuller (ADF) test, which has a significance level of $p<0.001$, confirmed that the fitting

analysis was stable on the training sample of influenza data used in this study. Because of this, the SARIMA model's parameters d and D are set to 0. Also, from what we saw in **Figure 2**, we can figure out that the SARIMA model's parameter S should be 12. After that, we found the values of the other factors p, q, P, and J by looking at the ACF and PACF plots (**Figure 3**). Furthermore, we utilized the grid search method for tuning the hyperparameters to identify the most optimal SARIMA model with the lowest AIC number. With an AIC value of 832.963 and a BIC value of 841.340, SARIMA(1, 0, 0)(1, 0, 0) (Song, 2017) is the best SARIMA model. The models' predictions are visually presented in **Figure 4**. **Figure 5** illustrates the comparison between the actual influenza case data and the predictions generated by the SARIMA model. The SARIMA model effectively depicts the seasonal variations and overall trends in monthly influenza cases. The predicted values and the real data exhibit discrepancies, particularly at peaks and troughs, which is expected in time series forecasting. Despite some occasional deviations, the SARIMA model's predictions adhere to the general pattern of the actual data, indicating its ability to accurately predict monthly influenza case trends.

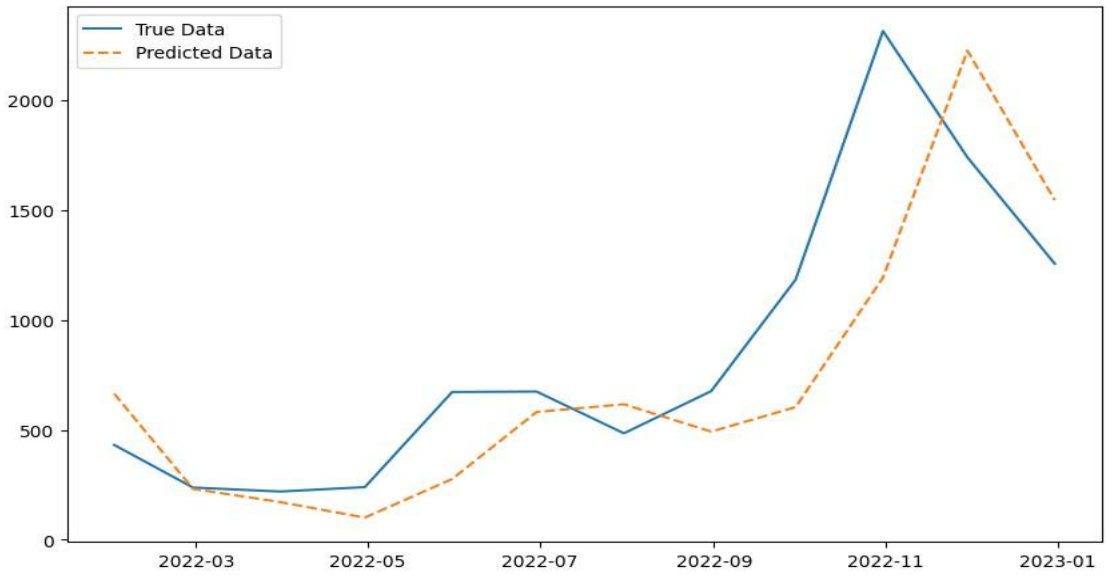

Figure 5. Comparative Evaluation of Monthly Influenza Cases Versus SARIMA Predictions

Forecasting Influenza Monthly Cases Applying the XGBoost Model

Applying the XGBoost model for monthly influenza case prediction requires meticulous selection of the pertinent hyperparameters, which is of utmost importance. We used grid search and five-fold cross-validation to identify the optimal combination of hyperparameters. The XGBoost hyperparameters that yielded the best results were: learning_rate=0.1, max_depth=3, n_estimators=100. Subsequently, we used the highly efficient XGBoost model to train on the training dataset.

Figure 6 shows how real monthly influenza incidence relates to the XGBoost model's projections. The projections fairly reflect the overall trend and the monthly influenza case count variations throughout many seasons. The projections show a significant alignment by closely matching the data, particularly in rising and declining trends. The model shows little difference between the anticipated values and the actual data, particularly at the highest and lowest locations. Still, the model shows really good performance overall. Especially in significant changes, the XGBoost model faithfully captures the swings in monthly influenza incidence. While the model generally aligns with the overall pattern of the actual data, some differences are most noticeable at the peak and subsequent drop. The chart highlights the XGBoost model's proficiency in identifying patterns and areas where further improvement might minimize differences. It demonstrates the accuracy of the model in forecasting influenza cases accurately.

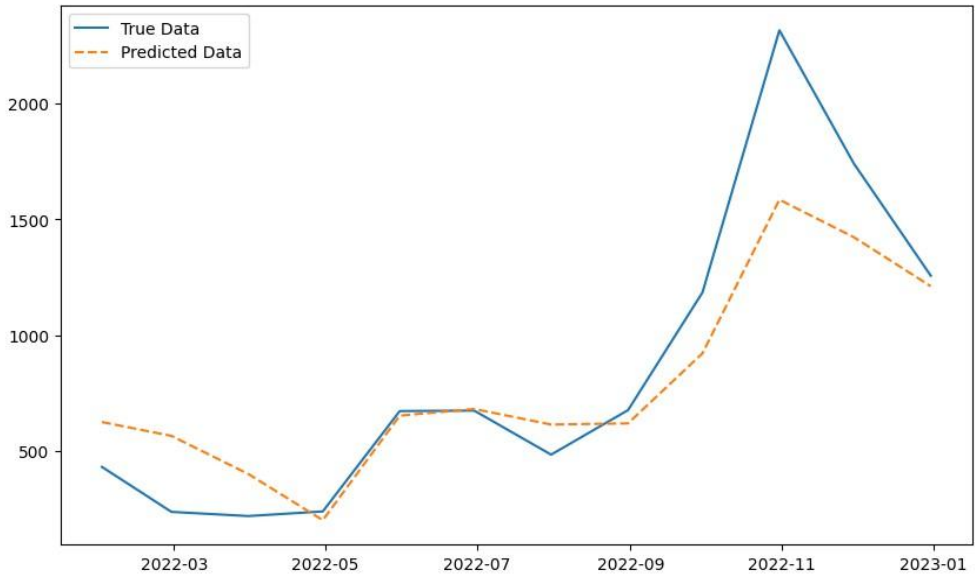

Figure 6. Comparative Evaluation of Monthly Seasonal Influenza Cases Versus XGBoost Predictions

Models Comparison

This study applied the XGBoost model to predict monthly influenza cases in Saudi Arabia by comparing the efficiency of the results with those obtained from the ARIMA and SARIMA models. We used four performance metrics to determine the most accurate and appropriate model. These are MSE, MAE, RMSE, and R^2. The results showed, as shown in **Table 1**, that the ARIMA model was the least efficient in predicting influenza cases, giving a mean squared error (MSE) of 43791.75, a mean absolute error (MAE) of 172.55, a root mean squared error (RMSE) of 209.26, and an R-squared (R^2) value of 0.775 for the training data. Meanwhile, the XGBoost model's results showed the highest performance, with an MSE of 3.75, an MAE of 1.39, an RMSE of 1.94, and an R^2 of 0.999. These results indicate that the XGBoost model outperforms other models in terms of accuracy, data suitability, and prediction of monthly influenza cases in Saudi Arabia.

Table 1. The Evaluation Matrix Compares the Performance of Models on Training and Test Sets for Monthly Seasonal Influenza Cases

	Train				Test			
Model	**MAE**	**MSE**	**RMSE**	**R2**	**MAE**	**MAE**	**MAE**	**MAE**
ARIMA(3,0,2)	172.5	43791.7	209.26	0.77	298.18	221664.44	470.81	0.44
ARIMA(1,0,0)(1,0,0)	165.09	55249.4	235.05	0.71	309.11	184091.04	429.05	0.53
XGBoost	1.38	3.747	1.93	0.99	192.61	75528.46	274.82	0.80

CONCLUSION

In this paper, out of all the machine learning models, the XGBoost method has a higher accuracy in the prognosis of pandemics compared to the ARIMA and SARIMA. It is established that the XGBoost results provide better prediction of the monthly flu cases than the ARIMA and SARIMA models. Since the obtained R^2 was the highest and MAE, MSE, and RMSE were the lowest among all the training and test data, more helpful for public health practice. This is because it is possible to improve the identification of flue cases, especially using XGBoost which gives more accurate estimates and that is developed under machine learning models. Due to such models, the public health initiatives are quicker and more effective in their response.

Since this study found that the XGBoost model is better for the monthly flu case prediction, it is hoped that the future researchers will look into the comparative performance of stochastic and fractional modeling for monitoring disease memories in order to know more about how diseases spread in the conditions of uncertainty (World Health Organization, 2023; Ali et al., 2024a; EL Guma et al., 2024; Alharbi et al., 2024; Almutairi et al., 2023; Ali et al., 2024b). These sophisticated methods, combined with the accuracy of the XGBoost model, can provide us with more extensive tools for predicting the spread of fatal diseases and simultaneously assist public

health officials in developing more effective methods to prevent disease attacks from occurring, thereby sparing lives and money.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Alharbi, S. A., Abdoon, M. A., Saadeh, R., Alsemiry, R. D., Allogmany, R., Berir, M., & EL Guma, F. (2024). Modeling and analysis of visceral leishmaniasis dynamics using fractional - order operators: A comparative study. *Mathematical Methods in the Applied Sciences*, *47*(12), 9918–9937. doi:10.1002/mma.10101

Ali, M., Alzahrani, S. M., Saadeh, R., Abdoon, M. A., Qazza, A., Al-kuleab, N., & EL Guma, F. (2024a). Modeling COVID-19 spread and non-pharmaceutical interventions in South Africa: A stochastic approach. *Scientific African*, *24*, e02155. doi:10.1016/j.sciaf.2024.e02155

Ali, M., Guma, F. E., Qazza, A., Saadeh, R., Alsubaie, N. E., Althubyani, M., & Abdoon, M. A. (2024b). Stochastic modeling of influenza transmission: Insights into disease dynamics and epidemic management. Partial Differential Equations in Applied Mathematics, 100886.

Aljandali, A. (2017). The Box-Jenkins methodology. In *Multivariate Methods and Forecasting with IBM® SPSS® Statistics*. Cham, Switzerland: Springer.

Almutairi, D. K., Abdoon, M. A., Salih, S. Y. M., Elsamani, S. A., Guma, F. E., & Berir, M. (2023). Modeling and analysis of a fractional visceral leishmaniosis with Caputo and Caputo–Fabrizio derivatives. *Journal of the Nigerian Society of Physical Sciences*, 1453-1453. doi:10.46481/jnsps.2023.1453

Alsobhi, A. (2022). Prediction of COVID-19 disease by ARIMA model and tuning hyperparameter through GridSearchCV. *Emerging Technologies in Data Mining and Information Security*, 543–551. doi:10.1007/979814051_54

Alsubaie, N., EL Guma, F., Boulehmi, K., Al-kuleab, N., & Abdoon, M. A. (2024). Improving influenza epidemiological models under Caputo fractional-order calculus. *Symmetry*, *16*(7), 929. doi:10.3390/sym16070929

Alzahrani, S. M., Saadeh, R., Abdoon, M. A., Qazza, A., Guma, F. E., & Berir, M. (2024). Numerical simulation of an influenza epidemic: Prediction with fractional SEIR and the ARIMA model. *Applied Mathematics & Information Sciences*, *18*(1), 1-12. doi:10.18576/amis/180101

Anderson, O. D. (1977). The Box-Jenkins approach to time series analysis. *RAIRO-Operations Research*, *11*(1), 29.

ArunKumar, K. E., Kalaga, D. V., Kumar, C. M. S., Chilkoor, G., Kawaji, M., & Brenza, T. M. (2021). Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA). *Applied Soft Computing*, *103*, 107161.

Arwaekaji, M., Sillabutra, J., Viwatwongkasem, C., & Soontornpipit, P. (2022). Forecasting influenza incidence in public health region 8 Udonthani, Thailand by SARIMA model. *Current Applied Science and Technology*, *22*(4). doi:10.55003/cast.2022.04.22.015

Badar, N., Ikram, A., Salman, M., Saeed, S., Mirza, H. A., Ahad, A., . . . Farooq, U. (2024). Evolutionary analysis of seasonal influenza A viruses in Pakistan 2020–2023. *Influenza and Other Respiratory Viruses*, *18*(2). doi:10.1111/irv.13262

Bezerra, A. K. L., & Santos, É. M. C. (2020). Prediction of the daily number of confirmed cases of COVID-19 in Sudan with ARIMA and Holt-Winters exponential smoothing. *International Journal of Development Research*, *10*(08), 394039413.

Box, G. (2013). Box and Jenkins: Time series analysis, forecasting and control. In *A Very British Affair: Six Britons and the Development of Time Series Analysis During the 20th Century* (p. 16215). London, UK: Palgrave Macmillan.

Chen, Q., Zheng, X., Shi, H., Zhou, Q., Hu, H., Sun, M., . . . Zhang, X. (2024). Prediction of influenza outbreaks in Fuzhou, China: Comparative analysis of forecasting models. *BMC Public Health*, *24*(1). doi:10.1186/s1288021858x

Chen, Y., Leng, K., Lu, Y., Wen, L., Qi, Y., Gao, W., ... & Dong, J. (2020). Epidemiological features and time-series analysis of influenza incidence in urban and rural areas of Shenyang, China, 2010–2018. *Epidemiology & Infection*, *148*, e29. doi:10.1017/S0950268820000151

Dancer, D., & Tremayne, A. (2005). R-squared and prediction in regression with ordered quantitative response. *Journal of Applied Statistics*, *32*(5), 483–493. doi:10.1080/02664760500079423

Dandachi, I., Alrezaihi, A., Amin, D., AlRagi, N., Alhatlani, B., Binjomah, A., . . . Aljabr, W. (2024). Molecular

surveillance of influenza A virus in Saudi Arabia: Whole-genome sequencing and metagenomic approaches. *Microbiology Spectrum*, *12*(8). doi:10.1128/spectrum.006624

Devlin, R. K. (2008). The influenza virus. In J. K. Silver (Ed.), *Influenza* (pp. 1–20). doi:10.5040/9798400670053

EL Guma, F. (2024). Comparative analysis of time series prediction models for visceral leishmaniasis: based on SARIMA and LSTM. *Applied Mathematics & Information Sciences*, *18*(1), 125-132. doi:10.18576/amis/180113

EL Guma, F., Abdoon, M. A., Qazza, A., Saadeh, R., Arishi, M. A., & Degoot, A. M. (2024). Analyzing the impact of control strategies on visceral leishmaniasis: A mathematical modeling perspective. *European Journal of Pure and Applied Mathematics*, *17*(2), 1213–1227. doi:10.29020/nybg.ejpam.v17i2.5121

EL Guma, F., Musa, A. G. M., Alkhathami, F. D., Saadeh, R., & Qazza, A. (2023). Prediction of visceral leishmaniasis incidences utilizing machine learning techniques. In *2023 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)* (pp. 1-6). Zarqa, Jordan: IEEE.

Hoque, K. E., & Aljamaan, H. (2021). Impact of hyperparameter tuning on machine learning models in stock price forecasting. *IEEE Access*, *9*, 163815–163830. doi:10.1109/access.2021.3134138

Kaur, J., Parmar, K. S., & Singh, S. (2023). Autoregressive models in environmental forecasting time series: A theoretical and application review. *Environmental Science and Pollution Research*, 30(8), 19617–19641. doi:10.1007/s11350225149

Khan, D. R., Patankar, A. B., & Khan, A. (2024). An experimental comparison of classic statistical techniques on univariate time series forecasting. *Procedia Computer Science*, *235*, 2730–2740. doi:10.1016/j.procs.2024.04.257

Kumar, D. S., Thiruvarangan, B. C., Vishnu, A., Devi, A. S., & Kavitha, D. (2022). Analysis and prediction of stock price using hybridization of SARIMA and XGBoost. In *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)* (pp. 1-4). Chennai, India: IEEE.

Kuran, F., Tanırcan, G., & Pashaei, E. (2023). Performance evaluation of machine learning techniques in predicting cumulative absolute velocity. *Soil Dynamics and Earthquake Engineering*, *174*, 108175. doi:10.1016/j.soildyn.2023.108175

Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. *Frontiers in Genetics*, *10*, 1077. doi:10.3389/fgene.2019.01077

Luo, J., Zhang, Z., Fu, Y., & Rao, F. (2021). Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. *Results in Physics*, *27*, 104462. doi:10.1016/j.rinp.2021.104462

Lv, C. X., An, S. Y., Qiao, B. J., & Wu, W. (2021). Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. *BMC Infectious Diseases*, *21*(1). doi:10.1186/s1287020650y

Man, H., Huang, H., Qin, Z., & Li, Z. (2023). Analysis of a SARIMA-XGBoost model for hand, foot, and mouth disease in Xinjiang, China. *Epidemiology and Infection*, *151*. doi:10.1017/s0950268823001905

Mills, T. C. (2019). ARIMA models for nonstationary time series. In *Applied Time Series Analysis* (pp. 57–69). doi:10.1016/b970-1813116.00001

Nelson, B. K. (1998). Time series analysis using autoregressive integrated moving average (ARIMA) models. *Academic Emergency Medicine*, *5*(7), 739–744. doi:10.1111/j.1552712.1998.tb02493.x

Nelson, M. I., & Holmes, E. C. (2007). The evolution of epidemic influenza. *Nature Reviews Genetics*, *8*(3), 196–205. doi:10.1038/nrg2053

Peixeiro, M. (2022). *Time series forecasting in Python*. Shelter Island, NY: Simon and Schuster.

Song, H. (2017, May 21). Review of Time Series Analysis and Its Applications With R Examples (3rd Edition) [Review of the book *Time Series Analysis and Its Applications With R Examples (3rd Edition)*, by R. H. Shumway & D. S. Stoffer]. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 800–802. doi:10.1080/10705511.2017.1299578

Sroka, Ł. (2024). Simulation analysis of artificial neural network and XGBoost algorithms in time series forecasting, Scientific Papers of Silesian University of Technology Organization and Management Series, *2024*(195). doi:10.29119/1643466.2024.195.34

Tenepalli, D., & TM, N. (2024). A systematic review on IoT and machine learning algorithms in e-healthcare. *International Journal of Computing and Digital Systems*, *16*(1), 27294.

World Health Organization. (2023). Global Influenza Surveillance and Response System (GISRS). Retrieved from

https://www.who.int/initiatives/global-influenza-surveillance-and-response-system

Yasmin, S., & Moniruzzaman, M. (2024). Forecasting of area, production, and yield of jute in Bangladesh using Box-Jenkins ARIMA model. *Journal of Agriculture and Food Research*, *16*, 101203.

Yenilmez, İ., & Mugenzi, F. (2023). Estimation of conventional and innovative models for Rwanda's GDP per capita: A comparative analysis of artificial neural networks and Box–Jenkins methodologies. *Scientific African*, *22*, e01902.

Zhang, L., Bian, W., Qu, W., Tuo, L., & Wang, Y. (2021). Time series forecast of sales volume based on XGBoost. *Journal of Physics: Conference Series*, *1873*(1), 012067. doi:10.1088/1746596/1873/1/012067

Zhao, Z., Zhai, M., Li, G., Gao, X., Song, W., Wang, X., . . . Qiu, L. (2023). Study on the prediction effect of a combined model of SARIMA and LSTM based on SSA for influenza in Shanxi Province, China. *BMC Infectious Diseases*, *23*(1), 71.